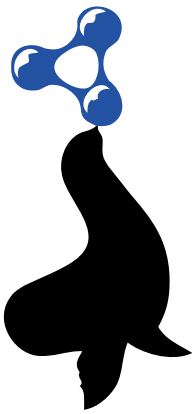
Design of the first evaluation campaign

Stuart Wrigley, Raúl García Castro, Liliana Cabral, Dorothee Reinhard, Cassia Trojahn dos Santos, Stephan Grimm, Mikalai Yatskevich



SEALS

Semantic Evaluation at Large Scale

FP7 – 238975

D3.2 Design of the First Evaluation Campaign

Coordinator: Stuart N. Wrigley
With contributions from: Stuart N. Wrigley, Raúl García-Castro, Liliana Cabral, Dorothee Reinhard, Cássia Trojahn dos Santos, Stephan Grimm, Mikalai Yatskevich
Quality Controller: Cássia Trojahn dos Santos
Quality Assurance Coordinator: Raúl García-Castro

Document Identifier:	SEALS/2010/D3.2/V1.0
Class Deliverable:	SEALS EU-IST-2009-238975
Version:	version 1.0
Date:	February 10, 2010
State:	final
Distribution:	public

Executive Summary

This deliverable summarises the work to date in the implementation of the SEALS methodology and design recommendations described in SEALS Deliverable 3.1 (García-Castro and Martín-Recuerda, 2009).

SEALS Deliverable D3.1 described the sequence of activities necessary to conduct the evaluation campaigns. This sequence was divided into four phases called *Initiation*, *Involvement*, *Preparation and Execution*, and *Dissemination*.

This deliverable covers the initial preparation of the first SEALS Evaluation Campaign. Chapters 3 and 4 describe the tasks performed during the *Initiation* and *Involvement* phases respectively.

Chapter 6 we summarises the plan of work required in the *Preparation and execution* and *Dissemination* phases together with two timelines.

The deliverable also contains a number of appendices which contain two forms of announcements for the first SEALS Evaluation Campaign as well as the specific evaluation scenarios which each of the five technology areas will be addressing. Finally, the general SEALS Evaluation Campaign terms and conditions are provided.

Document Information

IST Project Number	FP7 – 238975	**Acronym**	SEALS
Full Title	Semantic Evaluation at Large Scale		
Project URL	`http://www.seals-project.eu/`		
Document URL			
EU Project Officer	Carmela Asero		

Deliverable	**Number**	3.2	**Title**	Design of the First Evaluation Campaign
Work Package	**Number**	3	**Title**	Evaluation Campaigns and Semantic Technology Roadmaps

Date of Delivery	**Contractual**	M9	**Actual**	10-02-10
Status	version 1.0		final ☒	
Nature	prototype ☐ report ☒ dissemination ☐			
Dissemination level	public ☒ consortium ☐			

Authors (Partner)	Stuart N. Wrigley (USFD), Raúl García-Castro (UPM), Liliana Cabral (OU), Dorothee Reinhard (UZH), Cássia Trojahn dos Santos (INRIA), Stephan Grimm (FZI), Mikalai Yatskevich (OXF)			
Resp. Author	**Name**	Stuart N. Wrigley	**E-mail**	s.wrigley@dcs.shef.ac.uk
	Partner	University of Sheffield	**Phone**	+44 (114) 222 1880

Abstract (for dissemination)	This deliverable is concerned with the implementation of the evaluation campaign based upon the methodology and design recommendations made in SEALS Deliverable D3.1 (García-Castro and Martín-Recuerda, 2009). This deliverable covers the initial preparation of the first SEALS Evaluation Campaign and describes the tasks that have been performed during the *Initiation* and *Involvement* phases. Furthermore, the deliverable describes the steps to be taken over the next few months and the actors who are responsible for those steps.
Keywords	evaluation campaign, methodology, design, guidelines, timeline

Version Log			
Issue Date	**Rev No.**	**Author**	**Change**
11-01-2010	1	Stuart Wrigley	Creation of report outline
20-01-2010	2	Raúl García-Castro	Modifications in the structure and content
25-01-2010	3	Stuart Wrigley	Insertion of campaign scenarios
25-01-2010	4	Stuart Wrigley	Insertion of campaign announcements
25-01-2010	5	Stuart Wrigley	Insertion of campaign agreements from D3.1 as appendix
04-02-2010	6	Raúl García-Castro	Minor modifications
08-02-2010	7	Stuart Wrigley	Minor modifications
10-02-2010	8	Stuart Wrigley	Modifications as recommended by QC

Project Consortium Information

Participant's name	Partner	Contact
Universidad Politécnica de Madrid		Asunción Gómez-Pérez Email: asun@fi.upm.es
University of Sheffield		Fabio Ciravegna Email: fabio@dcs.shef.ac.uk
Forschungszentrum Informatik an der Universität Karlsruhe		Rudi Studer Email: studer@fzi.de
University of Innsbruck		Barry Norton Email: barry.norton@sti2.at
Institut National de Recherche en Informatique et en Automatique		Jérôme Euzenat Email: Jerome.Euzenat@inrialpes.fr
University of Mannheim		Heiner Stuckenschmidt Email: heiner@informatik.uni-mannheim.de
University of Zurich		Abraham Bernstein Email: bernstein@ifi.uzh.ch
Open University		John Domingue Email: j.b.domingue@open.ac.uk
Semantic Technology Institute International		Alexander Wahler Email: alexander.wahler@sti2.org
University of Oxford		Ian Horrocks Email: ian.horrocks@comlab.oxford.ac.uk

Table of Contents

List of figures 7

List of tables 8

1 Introduction 9

2 Glossary of terms 11
2.1 Glossary for evaluation campaigns . . . 11
2.2 Glossary for evaluations . . . 11

3 Tasks performed in the *Initiation* phase 12
3.1 Identify organizers . . . 12
3.1.1 Membership . . . 13
3.2 Organize evaluation scenarios . . . 13

4 Tasks performed in the *Involvement* phase 15
4.1 Announce the evaluation campaign . . . 15
4.2 Provide registration mechanisms . . . 16

5 Community engagement 17
5.1 Incorporating external persons in SEALS . . . 18
5.2 Promotion of SEALS and the SEALS Evaluation Campaign . . . 18

6 Future activities and timeline 20
6.1 Tasks for the next period . . . 20
6.1.1 *Preparation and execution* phase . . . 20
6.1.2 *Dissemination* phase . . . 21
6.2 Timeline . . . 22
6.2.1 Participants . . . 22
6.2.2 SEALS Partners . . . 22

7 Conclusions 25

References 25

A Evaluation scenarios 27
A.1 WP10: Ontology Engineering Tools . . . 27
A.2 WP11: Storage and Reasoning Systems . . . 27
A.3 WP12: Matching Tools . . . 27
A.4 WP13: Semantic Search Tools . . . 28
A.4.1 Automated Phase . . . 28
A.4.2 User-in-the-loop Phase . . . 28
A.5 WP14: Semantic Web Service Tools . . . 29

B General announcement of the SEALS Evaluation Campaigns 30

C Long announcements of evaluation campaigns 32
C.1 Common to all tool areas . . . 32
C.1.1 Instructions on how to participate . . . 32
C.1.2 Evaluation Process . . . 32
C.1.3 How to find out more . . . 33
C.2 Ontology Engineering Tools . . . 33
C.3 Storage and Reasoning Systems . . . 34
C.4 Matching Tools . . . 35
C.4.1 Evaluation Description . . . 35
C.4.2 Evaluation Test Data . . . 35
C.5 Semantic Search Tools . . . 36
C.5.1 Introduction . . . 36
C.5.2 Criteria . . . 36
C.5.3 Two phase approach . . . 37
C.5.4 Evaluation Test Data . . . 37
C.6 Semantic Web Service Tools . . . 37

D Evaluation campaign agreements 39
D.1 Terms of participation . . . 39
D.2 Use rights . . . 39
D.2.1 Rule applying to the non-public use of the data . . . 39
D.2.2 Rules applying to evaluation campaign participants . . . 39
D.2.3 Rules applying to people who did not participate in an evaluation campaign . . . 40
D.2.4 Rules applying to other cases . . . 40

List of Figures

1.1 The evaluation campaign process. 9

3.1 *Initiation* phase of the evaluation campaign process. 12

4.1 *Involvement* phase of the evaluation campaign process. 15

6.1 *Preparation and execution* phase of the evaluation campaign process. . 21
6.2 *Dissemination* phase of the evaluation campaign process. 22
6.3 Timeline of events involved in the first SEALS Evaluation Campaign with tasks related to the evaluation participants shown above the timeline and tasks related to SEALS partners shown below the timeline. . . 24

List of Tables

3.1 Membership of the E.C.O.C.. 13
3.2 Membership of each E.C.E.C.. 14

1. Introduction

The SEALS project will create a lasting reference infrastructure for semantic technology evaluation (the SEALS Platform) and thus facilitate the continuous evaluation of semantic technologies at a large scale. The SEALS Platform will be an independent, open, scalable, extensible and sustainable infrastructure that will allow the evaluation of semantic technologies by providing an integrated set of evaluation services and test suites.

The SEALS project will take place in two 18-month stages and in each of these stages different evaluation campaigns will be performed for each of the technologies covered in the project. The SEALS Platform will be used in these evaluation campaigns and their results will be employed in creating semantic technology roadmaps that will identify sets of efficient and compatible tools for developing large-scale semantic applications.

This document focusses on the design of the first SEALS Evaluation Campaign and draws heavily upon the methodology and design recommendations made in SEALS Deliverable D3.1 (García-Castro and Martín-Recuerda, 2009). SEALS Deliverable D3.1 described the sequence of activities necessary to conduct the evaluation campaigns. This sequence, shown in Figure 1.1, was divided into four phases called *Initiation*, *Involvement*, *Preparation and Execution*, and *Dissemination*.

Figure 1.1: The evaluation campaign process.

The first SEALS Evaluation Campaign, which will address five core semantic technology areas, will run until November 2010. This deliverable covers the initial preparation of the first SEALS Evaluation Campaign and starts describing the tasks performed during the *Initiation* and *Involvement* phases in chapters 3 and 4, respectively. Specifically, these chapters address the formation of two types of committee: the *Evaluation Campaign Organizing Committee* (E.C.O.C.) and a number of *Evaluation Campaign Executing Committees* (E.C.E.C.s) which will oversee the running of the first (and future) SEALS Evaluation Campaigns and specify the individual memberships of those committees. Furthermore, Chapter 4 describes the work undertaken to date by the E.C.E.C.s.

As can be seen from Figure 1.1, the *Preparation and Execution* phase runs concurrently with the *Involvement* phase. In Chapter 6 we summarise the plan of work required in the *Preparation and execution* and *Dissemination* phases together with two timelines. Each timeline targets a different set of people associated with SEALS: the set of participants and the members of the SEALS consortium.

Chapter 5 describes a number of activities related to community engagement. This chapter addresses the role of external persons both in terms of their participation in the

E.C.O.C. and E.C.E.C.s as well as their potential participation in the wider research of the SEALS project. Chapter 5 also summarises efforts to promote SEALS and the first SEALS Evaluation Campaign.

The deliverable also contains a number of appendices which contain two forms of announcements for the first SEALS Evaluation Campaign (as described in Section 4.1) as well as the specific evaluation scenarios which each of the five technology areas will be addressing. Finally, the general SEALS Evaluation Campaign terms and conditions are included in Appendix D.

2. Glossary of terms

This chapter presents a glossary of terms, extracted from (García-Castro and Martín-Recuerda, 2009), that are grouped in two main groups: terms related to evaluation campaigns and terms related to evaluations.

2.1 Glossary for evaluation campaigns

Evaluation campaign. An evaluation campaign is an activity where one or several evaluation campaign scenarios are performed over several tools. In SEALS there will be two evaluation campaigns for each of the types of semantic technologies.

Evaluation campaign scenario (or evaluation scenario). An evaluation campaign scenario is an evaluation where several tools are evaluated and compared according to common test data.

Evaluation campaign participant. An evaluation campaign participant is a person or organization that participates with a tool in one or several evaluation campaign scenarios.

2.2 Glossary for evaluations

Evaluation target (or tool). An evaluation target is a specific tool to be evaluated in SEALS that fits one of the semantic technology categories.

Evaluation use case. An evaluation use case is the systematic examination of the behaviour of a tool under certain use conditions and with certain input data. An evaluation use case is composed of an evaluation description, an evaluation workflow and evaluation test data.

Evaluation criterion. An evaluation criterion determines which characteristic of a certain type of tool is assessed in an evaluation use case. Several evaluation criteria can be used in a certain evaluation use case.

Evaluation metric. An evaluation metric is an instrument for measuring the outcome of an evaluation use case. Evaluation metrics are related to the evaluation criteria.

Evaluation description. An evaluation description is the documentation needed to understand and replicate a certain evaluation use case.

Evaluation workflow. An evaluation workflow determines the way in which an evaluation use case is conducted in terms of its input, output and operations, plus the way in which the results are interpreted using evaluation metrics.

Evaluation test data. Evaluation test data are input data that can be used in one or several evaluations use cases.

3. Tasks performed in the *Initiation* phase

The Initiation phase comprises the set of tasks where the different people involved in the organization of the evaluation campaign and the evaluation scenarios are identified and where the different evaluation scenarios are defined. These tasks and their interdependencies are shown in Figure 3.1.

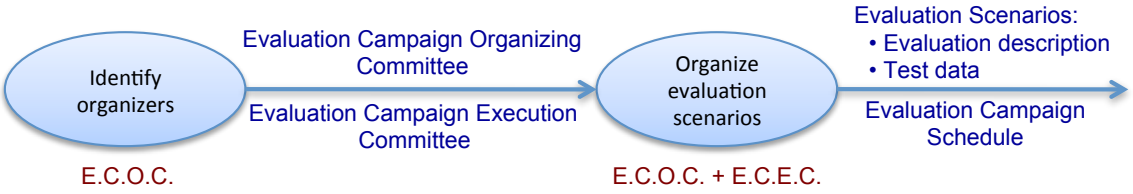


Figure 3.1: *Initiation* phase of the evaluation campaign process.

The first aspect of this task is the identification of the organisers; this is concerned with the creation of the two committees, described in Section 3.1, namely the E.C.O.C. and E.C.E.C.. The second aspect is the organisation of the evaluation scenarios, which is described in Section 3.2.

3.1 Identify organizers

The tasks of the evaluation campaign process are carried out by different actors according to the kind of roles that must be performed in each task. The SEALS Deliverable D3.1 (García-Castro and Martín-Recuerda, 2009) defined a number of different actors who will participate in the evaluation campaign process. This section presents the different kinds of actors involved in such process.

- **Evaluation Campaign Organizing Committee** (E.C.O.C.). The E.C.O.C. is in charge of the general organization and monitoring of the evaluation campaign. In SEALS there is one E.C.O.C. for all the evaluation campaigns and is composed of the SEALS Executive Project Management Board, the SEALS research work package leaders and other prominent external people.
- **Evaluation Campaign Executing Committee** (E.C.E.C.). The E.C.E.C. is in charge of organizing the evaluation scenarios that are performed in the evaluation campaign and of taking them to a successful end. In SEALS there will be at least one E.C.E.C. for each technology area (e.g., one E.C.E.C. for semantic search tool evaluation, one E.C.E.C. for matching tool evaluation, etc.).
- **Participants**. The evaluation campaign participants are tool providers or people with the permission of tool providers that participate with a tool in the evaluation campaign.

For the first evaluation campaign (v1) membership of the E.C.O.C. and E.C.E.C. consists mainly of SEALS partners (see also Chapter 5). Therefore, membership of the E.C.O.C. and E.C.E.C. will be:

- **Evaluation Campaign Organizing Committee** (E.C.O.C.). SEALS Executive Project Management Board (EPMB) and the SEALS research work package leaders.
- **Evaluation Campaign Executing Committee** (E.C.E.C.). A subset of the relevant research work package as defined by the work package leader.

3.1.1 Membership

E.C.O.C.

The E.C.O.C. is composed of the SEALS Executive Project Management Board, the SEALS work package leaders and other prominent external people. The constituent members are shown in Table 3.1. Note that the E.C.O.C. also contains the WP3 Leader (coordinator of the evaluation campaign organisation work package) and the WP2 Leader (coordinator of the dissemination and community building work package). Furthermore, as specified in the Description of Work, Fabio Ciravegna (USFD) has the role of *Evaluation Campaigns Coordinator* and therefore will act as the chair of the E.C.O.C. with Stuart Wrigley (USFD; WP3 Leader) acting as his deputy.

Table 3.1: Membership of the E.C.O.C..

Member	**Affiliation**	**Project Role**
Asunción Gómez-Pérez	UPM	EPMB Member
Fabio Ciravegna	USFD	EPMB Member
Lyndon Nixon	STI2	WP2 Leader
Stephan Grimm	FZI	WP10 Leader
Giorgos Stoilos	OXF	WP11 Leader
Jérôme Euzenat	INRIA	WP12 Leader and EPMB Member
Stuart Wrigley	USFD	WP13 Leader and WP3 Leader
Liliana Cabral	OU	WP14 Leader

E.C.E.C.

Each E.C.E.C. contains a subset of the relevant WP personnel as shown in Table 3.2.

3.2 Organize evaluation scenarios

The E.C.E.C. is in charge of organizing the evaluation scenarios that are performed in the evaluation campaign and of taking them to a successful end. The evaluation scenarios identified by each E.C.E.C. are described in Appendix A.

Table 3.2: Membership of each E.C.E.C..

WP	**WP Name**	**Members**	**Affiliation**
10	Ontology Engineering Tools	Stephan Grimm Raúl García-Castro	FZI UPM
11	Storage and Reasoning Systems	Mikalai Yatskevich	OXF
12	Matching Tools	Cassia Trojahn dos Santos Heiner Stuckenschmidt	INRIA UMA
13	Semantic Search Tools	Stuart Wrigley Dorothee Reinhard	USFD UZH
14	Semantic Web Service Tools (SWS Tool Discovery)	Liliana Cabral Mick Kerrigan	OU UIBK
14	Semantic Web Service Tools (Semantic Service Selection Contest)	Matthias Klusch (external person)	DFKI

4. Tasks performed in the *Involvement* phase

The *Involvement* phase comprises the set of tasks in which the evaluation campaign is announced and participants show their interest in participating by registering for the evaluation campaign. These tasks and their interdependencies are shown in Figure 4.1.

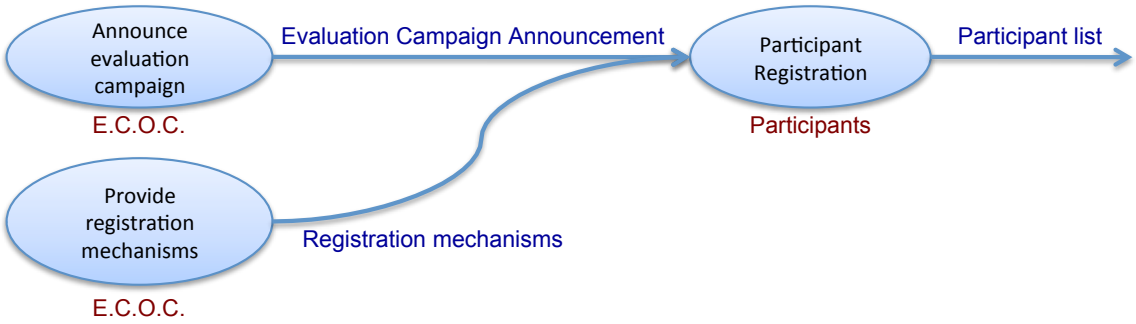


Figure 4.1: *Involvement* phase of the evaluation campaign process.

The tasks performed in this phase have been the preparation and dissemination of the evaluation campaign announcements and the development of mechanisms for participant registration, presented in Sections 4.1 and 4.2, respectively. By the time of writing this deliverable, the evaluation campaigns have been announced and participant registration has started.

4.1 Announce the evaluation campaign

As part of the *Involvement* phase, each E.C.E.C. has created two forms of public announcement: a 'short' and a 'long' announcement. These will be used both by WP2 and the E.C.E.C.s themselves to advertise the SEALS Evaluation Campaigns. Each type of announcements can be found as appendices to this deliverable (appendix B and C).

The short announcements have been combined to create a single announcement which gives a general description of the SEALS project and a broad introduction to each technology area's evaluation campaign. The target audience for this announcement is the semantic technology community in general and allied fields. Specifically, these distribution channels will be those maintained by STI International for such promotional activities as well as providers who have already registered an interest in SEALS via the SEALS Portal.

Each longer announcement will be issued independently and will be targeted at researchers and tool developers known to be active in that particular tool field. These announcements give detailed information regarding the goals and evaluation scenarios involved in the evaluation campaign as well as information regarding how to participate. These announcements will be sent to the vendors identified as part of SEALS T2.1 (described in D2.1 and enumerated on the SEALS private wiki) in which each

of the work packages responsible for a particular technology area created a list of potential participants.

4.2 Provide registration mechanisms

Each E.C.E.C. has also defined the information they wished to collect at registration regarding individual participants and general information regarding the tools that they are interested in evaluating. This information was drawn from the metadata that each tool area defined in D{10-14}.1 augmented by additional information identified to date.

This information was passed to WP2 which formed the basis of the design of the registration mechanisms put in place on the SEALS Portal.

The general terms and conditions associated with participation in the SEALS Evaluation Campaigns (see appendix D) have also been made available on the Portal and are explicitly referred to in each E.C.E.C.'s detailed announcement.

5. Community engagement

The SEALS Evaluation Campaigns are naturally community-oriented exercises with the overall goal of the project being to create a lasting infrastructure and best practice which persists after the funded duration of the project. At the core of the community involvement lies the participation of tool providers / developers who wish to evaluate their tools as well as tool adopters who will use the SEALS technology roadmaps and evaluation results to inform their technology decision processes.

However, it is also a goal of the SEALS project to involve technology providers and adopters in the design and execution of the SEALS Evaluation Campaigns themselves. This aspect is explicitly described in the methodology and design recommendations made in SEALS Deliverable D3.1 (García-Castro and Martín-Recuerda, 2009). In practice this relates to the involvement of persons external to the project in both the *Evaluation Campaign Organizing Committee* (E.C.O.C.) and each of the five technology areas' *Evaluation Campaign Executing Committees* (E.C.E.C.s). Such persons don't necessarily have to be drawn from technology developers or adopters, it is the intention of the SEALS consortium that high profile academic and industrial researchers also be involved.

However, for the first SEALS Evaluation Campaign, less emphasis has been placed on external membership of the E.C.O.C. and E.C.E.C.s. This has largely been due to pragmatic considerations regarding the schedule of work ahead of the first SEALS Evaluation Campaign. A campaign of this scale has not been attempted before in the semantic technologies field and many aspects were the subject of planned SEALS research activities. For instance, the methodology defining the procedures and actors required to execute the SEALS Evaluation Campaigns (D3.1; García-Castro and Martín-Recuerda (2009)) was only completed in late November 2009 by which stage, the design of each of the five technology area's evaluation campaigns had been completed. Furthermore, work is still ongoing regarding publicising SEALS and the two SEALS Evaluation Campaigns (see Section 5.2). It is through this publicity, the outcome of the first SEALS Evaluation Campaign and the association of SEALS with respected Semantic Web community events that will make involvement in the SEALS initiative attractive to persons external to SEALS.

Furthermore, it is hoped that existing evaluation efforts by members of the semantic community be aligned or even incorporated into the relevant SEALS Evaluation Campaigns. Indeed, the ability to incorporate external persons onto an E.C.E.C. provides a mechanism for this. In collaboration with the relevant E.C.E.C. and WP leader, the organiser can discuss the form of collaboration relevant to their evaluation effort. Indeed, such a collaboration has been put in place for the SEALS evaluation of Semantic Web Service Tools in which the Semantic Service Selection Contest (S3C) has been incorporated as an evaluation scenario (see Appendix A) and an E.C.E.C. within WP14 (see Table 3.2).

5.1 Incorporating external persons in SEALS

In addition to the inclusion of external representatives on the E.C.O.C. and the various E.C.E.C.s, the consortium has also put procedures in place for the creation of *Associated Partners*. Such persons could be identified by WP leaders on the basis that potential candidates can contribute concretely to the work done in SEALS, and specifically in their technology area. The procedure that has been accepted by the SEALS consortium is the following:

1. WP leaders can either identify potential candidates which can contribute concretely to the work done in SEALS or be approached by such candidates.
2. Potential candidates formulate a short justification for their inclusion as associated partners including which SEALS work package they plan to contribute. This must be with the agreement of the respective WP leader.
3. The proposal is made by e-mail to the Project Management Board (PMB).
4. A decision will be made within one week based on the responses of the PMB members.
5. If accepted, a formal notification will be send to the new Associated Partner and their details, including planned contribution to SEALS, will be published on the SEALS Portal.

Associated partners can be invited to (parts of) SEALS meetings and participate in SEALS activities. Throughout their involvement with SEALS, the respective WP leader will act as the main liaison with the associated partner.

It must be noted that the granting of Associated Partner status does not infer any entitlement to SEALS funding nor any formal place in the SEALS consortium.

5.2 Promotion of SEALS and the SEALS Evaluation Campaign

SEALS Deliverable D2.1 (Nixon, 2009a) described the SEALS community building and dissemination plan which is split into two phases covering the lifespan of the project (M1-18, and M19-36). It identified three target groups to be addressed by the plan: the research community, the tool provider community and technology adopters. Clearly, the first phase is largely focused on the first two groups (research and tool provider communities), with the third group (technology adopters) receiving more focus in the second phase once the first SEALS Evaluation Campaign results are available.

SEALS Deliverable D2.4 (Grimm et al., 2009) details the efforts made in the first six months of the SEALS project to raise the profile of the SEALS Evaluation Campaigns and to encourage involvement in the coming months. In brief, a number of publicity materials have been produced — a leaflet, a poster and a regular newsletter — which have been distributed or presented via various channels. For example, the poster and leaflets were distributed at a number of different events including European Semantic Web Conference (ESWC) 2009, Future Internet Symposium (FIS) 2009, International Semantic Web Conference (ISWC) 2009, and Asian Semantic Web Conference (ASWC) 2009.

The newsletter is intended to provide interested readers with more in-depth information about SEALS and its current activities. The first issue, which was distributed prior to ISWC 2009 (October 2009), had the goal of interesting people in SEALS and encourage them to become involved by joining the SEALS community; the latter facilitated by the creation of the SEALS Community Portal (Nixon, 2009b).

The SEALS project and the outcome of the first SEALS Evaluation Campaign will be publicised at the workshop planned to be held at International Semantic Web Conference (ISWC) 2010 in Shanghai, China. This is a prestigious and high profile conference which attracts delegates from throughout the international semantic technology community. As such, this is the ideal event at which the SEALS consortium can both publish the results and analyses of the first SEALS Evaluation Campaign but also involve persons external to the project in the planning and design of the second SEALS Evaluation Campaign.

6. Future activities and timeline

This chapter describes the tasks — as identified in SEALS Deliverable D3.1 (García-Castro and Martín-Recuerda, 2009) — that will form the focus of work related to the organisation of the first SEALS Evaluation Campaign. This chapter has been split into two parts: the first part addresses the two remaining phases and their associated tasks; the second part describes the timeline by which these tasks will be completed. Note that this latter part is further split into two to emphasise the differing responsibilities for SEALS Evaluation Campaign participants and SEALS Evaluation Campaign organisers (the E.C.E.C.s).

6.1 Tasks for the next period

This section summarises the tasks that need to be accomplished as part of the *Preparation and execution* and *Dissemination* phases. These tasks will largely be the responsibility of the individual E.C.E.C.s for each of the five technology areas. However, some of the tasks described also require input from the technology providers / participants.

6.1.1 *Preparation and execution* phase

The *Preparation and execution* phase comprises the set of tasks that must be performed to insert the participating tools into the evaluation infrastructure, to execute each of the evaluation scenarios, and to analyse the evaluation results. These tasks and their interdependencies, shown in Figure 6.1, are the following:

1. **Provide evaluation materials**.
   The E.C.E.C. must provide to the registered participants all the evaluation materials needed in the evaluation, including:
   (a) Instructions on how to participate.
   (b) The evaluation description.
   (c) The evaluation test data.
   (d) The evaluation infrastructure.
   (e) Any software needed for the evaluation.
2. **Insert tools**.
   Once the participants have all the evaluation materials, they must insert their tools into the evaluation infrastructure and ensure that these tools are ready for the evaluation execution.
3. **Perform evaluation**.
   The evaluation is executed over all the participating tools and the evaluation results of all the tools are collected.

4. **Analyse results**.
   Once the evaluation results of all the tools are collected, they are analysed both individually for each tool and globally including all the tools. This results analysis must be reviewed in order to get agreed conclusions. Therefore, if the results are analysed by the E.C.E.C. then this analysis must be reviewed by the participants and vice versa, that is, if the results are analysed by the participants they must be reviewed by the E.C.E.C..

Figure 6.1: *Preparation and execution* phase of the evaluation campaign process.

6.1.2 *Dissemination* phase

The *Dissemination* phase comprises the set of tasks that must be performed to disseminate the evaluation campaign results by publicly presenting them and to make all the evaluation campaign results and resources available. The tasks that compose this phase can be performed either independently for each evaluation scenario or covering all the evaluation scenarios in each task. These tasks and their interdependencies, shown in Figure 6.2, are the following:

1. **Present results**.
   The E.C.O.C., the E.C.E.C. and the participants will present and discuss the results of the evaluation campaign. It is envisaged that this will take the form of a workshop at ISWC 2010. The workshop will also be used to obtain feedback about the evaluation campaign.

2. **Publish results**.
   The E.C.O.C., the E.C.E.C. and the participants will publish the results of the evaluation campaign and of each of the tools either as workshop proceedings or journal special issues.

3. **Finalize**.
   All the evaluation resources used in the evaluation campaign will be made public. The final report — SEALS Deliverable D3.3 — will include the results of the campaign as well as recommendations for improving the campaign process for the second SEALS Evaluation Campaign.

Figure 6.2: *Dissemination* phase of the evaluation campaign process.

6.2 Timeline

This section provides two timelines describing the work planned until the completion of the first SEALS Evaluation Campaign. Each timeline targets a different set of people associated with SEALS: the set of participants and the members of the SEALS consortium. However, each timeline should not be read in isolation: they are inherently interdependent and only together do they give a full description of the project's future work. Both timelines are shown Figure 6.3 with more details in Sections 6.2.1 and 6.2.2.

6.2.1 Participants

- March - July 2010: Register for participation at the SEALS Portal
- June 2010: Obtain evaluation materials
- June - July 2010: Insert tools into the SEALS Platform
- August 2010: In conjunction with E.C.E.C., perform evaluation
- September 2010: Result analysis
- Early November 2010: Workshop at ISWC 2010

6.2.2 SEALS Partners

- February 2010: Evaluation Campaign Announced.
  - The E.C.O.C., led by the WP2 leader, will disseminate the general announcement including all evaluation campaigns to the existing distribution channels (mailing lists, blogs, SEALS Portal, etc.).
  - Each E.C.E.C. will send technology area-specific announcements to all the vendors identified for T2.1 and T2.5 announcing details of their evaluation campaigns including timelines aimed at the potential participants.
- Mid February 2010: Each E.C.E.C. sends WP2 leader requirements for participant registration. This will specify all information the E.C.E.C. wishes to be captured both about registrants and their tools.
- End February 2010: E.C.O.C., led by the WP2 leader, will provide registration mechanisms in the SEALS Portal.

- End March 2010: SEALS Repositories (data, tools, results) ready.
- End April 2010: SEALS Evaluation Description Repository ready.
- End April 2010: Each E.C.E.C. inserts the test data into the SEALS Platform.
- End April 2010: Tools may be inserted.
- End May 2010: Each E.C.E.C. inserts the evaluation scenarios (description + workflow) into the SEALS Platform.
- End May 2010: Each E.C.E.C. makes evaluation materials available. This will include the provision of any software needed for the evaluation, inserted/connected into/with the SEALS Platform.
- End May 2010: Basic Runtime Evaluation Service ready.
- End June 2010: Advnced Runtime Evaluation Service ready.
- End June 2010: Each E.C.E.C. inserts basic results analysis tools into the SEALS Platform.
- End July 2010: Each E.C.E.C. inserts advanced results analysis tools into the SEALS Platform.
- End July 2010: SEALS Platform available.
- August 2010: Evaluations conducted by each E.C.E.C. (possibly directly including participants, e.g., WP13).
- End August 2010: Evaluation results.
- September 2010: Result analysis.
- Early November 2010: Workshop at ISWC 2010.
- End November 2010: Each E.C.E.C. makes evaluation resources public.
- End November 2010: Each E.C.E.C. publishes results of evaluation.
- End November 2010: E.C.O.C. and E.C.E.C.s produce final report D3.3.

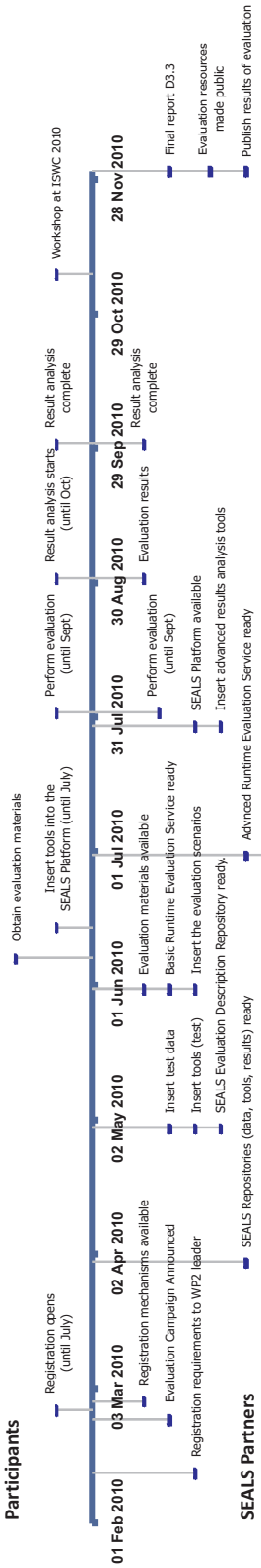


Figure 6.3: Timeline of events involved in the first SEALS Evaluation Campaign with tasks related to the evaluation participants shown above the timeline and tasks related to SEALS partners shown below the timeline.

7. Conclusions

This deliverable has summarised the work to date in the implementation of the SEALS methodology and design recommendations described in SEALS Deliverable 3.1 (García-Castro and Martín-Recuerda, 2009).

SEALS Deliverable D3.1 described the sequence of events necessary to conduct the evaluation campaigns. This sequence was divided into four phases called *Initiation*, *Involvement*, *Preparation and Execution*, and *Dissemination*.

The *Initiation* phase has been completed and the *Involvement* phase is nearing completion with only the formal announcements of the five technology area SEALS Evaluation Campaigns still outstanding (planned for late-February 2010).

The *Preparation and Execution* phase tuns concurrently with the *Involvement* phase. We have summarised the steps needed to be taken in order to take the first SEALS Evaluation Campaign to a successful completion. Indeed, work is well under way to fulfill the objectives identified in this phase (more details can be found in other SEALS Deliverables and Management Reports). Planning is underway regarding the *Dissemination* phase: for example, the SEALS Consortium expect to hold a workshop at International Semantic Web Conference (ISWC) 2010 in Shanghai, China since this is a prestigious and high profile conference which attracts delegates from throughout the international semantic technology community.

We have summarised the plan of work required in the *Preparation and execution* and *Dissemination* phases together with two timelines for accomplishing this work. Activities in the *Preparation and execution* phase address the provision of the evaluation materials (instructions, data, infrastructure, etc), the ability of participants to insert their tool(s) into the SEALS platform and subsequently have them benchmarked as part of the first SEALS Evaluation Campaign. The final core activity involves the analysis of the evaluation results. The *Dissemination* phase addresses the presentation and publication of these results and their associated analyses. A final project report will provide an overview of these results as well as gather feedback and recommendations which will inform the design of the second SEALS Evaluation Campaign. The timelines provide two different perspectives on the work planned over the next 9 months: one shows the schedule of activities for each SEALS Evaluation Campaign participant. The second gives more detail of the tasks to be completed by SEALS partners and in particular the E.C.O.C. and E.C.E.C.s.

References

R. García-Castro and F. Martín-Recuerda. D3.1: SEALS methodology for evaluation campaigns v1. Technical report, SEALS Consortium, 2009.

S. Grimm, L. Nixon, and R. García-Castro. D2.4: First report for community building and dissemination. Technical report, SEALS Consortium, 2009.

L. Nixon. D2.1: Plans for community building and dissemination. Technical report, SEALS Consortium, 2009a.

L. Nixon. D2.3: SEALS community portal. Technical report, SEALS Consortium, 2009b.

A. Evaluation scenarios

A.1 WP10: Ontology Engineering Tools

The following naming scheme has been adopted to enable easy identification of evaluation campaign scenarios:

`<ToolType> <Name/Criterion> <Year>`

The evaluation scenarios are as follows (OET = ontology engineering tool):

- OET Conformance 2010
  Testing conformance of OETs according to a language specification.
- OET Interoperability 2010
  Testing interoperability between OETs when using an interchange language.
- OET Scalability 2010
  Testing loading times and memory consumption for handling large ontologies with OETs.

A.2 WP11: Storage and Reasoning Systems

The following naming scheme has been adopted to enable easy identification of evaluation campaign scenarios:

`<ToolType> <Name/Criterion> <Year>`

The evaluation scenarios are as follows (DLBS = description logic based system):

- DLBS Classification 2010
- DLBS Class satisfiability 2010
- DLBS Ontology satisfiability 2010
- DLBS Logical entailment 2010

A.3 WP12: Matching Tools

The following naming scheme has been adopted to enable easy identification of evaluation campaign scenarios:

`<ToolType> <Name/Criterion> <Year>`

The evaluation scenarios are as follows (MT = Matching Tool):

- MT Benchmark 2010
  Criteria: conformance with expected results, efficiency in terms of memory consumption and execution time, and interoperability.

- MT Anatomy 2010
  Criteria: conformance with expected results, efficiency in terms of memory consumption and execution time, and interoperability.

- MT Conference 2010
  Criteria: conformance with expected results, efficiency in terms of memory consumption and execution time, interoperability, and alignment coherence.

A.4 WP13: Semantic Search Tools

The following naming scheme has been adopted to enable easy identification of evaluation campaign scenarios:

`<ToolType> <Phase> <Name/Criterion> <Year>`

The evaluation scenarios are listed below (SST = Semantic Search Tool). It should be noted that participants of the Semantic Search Evaluation Campaign must participate in all scenarios.

A.4.1 Automated Phase

1. SST Automated Search Performance 2010
   The tool's core search quality in terms of precision, recall, etc.

2. SST Automated Performance and Scalability 2010
   The tool will be tested for its ability to load, handle and perform queries on large data sets; scalability will be considered along three dimensions:

   - Average time to execute query with respect to ontology size.
   - Average CPU load to execute query with respect to ontology size.
   - Average memory to execute query with respect to ontology size.

3. SST Automated Quality of Documentation 2010
   Testing whether the natural language of the tool's documentation is easy to understand and well structured.

A.4.2 User-in-the-loop Phase

1. SST User Usability 2010

   - How do the end-users react to the tool's query language? Do they like the tool? Are they able to express their questions effectively and fast? Is the language easy to understand and learn? These aspects of user satisfaction will be assessed in the questionnaires.
   - Interpretation of the correlations between the user demographics and the measures.

2. SST User Query expressiveness 2010

    - Testing the formal (by asking the test subjects in the questionnaire), as well as, the practical expressiveness (by running queries in order to test the actual coverage and robustness) of the tool's query language, i.e. if the tool is able to answer complex queries.
    - Were all queries answered by the tool; i.e., could an answer be found?

A.5 WP14: Semantic Web Service Tools

The following naming scheme has been adopted to enable easy identification of evaluation campaign scenarios:

`<ToolType> <Name/Criterion> <Year>`

The evaluation scenarios are as follows (SWS = Semantic Web Services):

- SWS Tool Discovery Evaluation 2010
  Retrieval and execution time performance evaluation of SWS tools for the discovery of Web Services based on their semantic annotations.
  TC: WSMO and WSMO-Lite descriptions of OPOSSUM services (also in RO)
- SWS S3 (Semantic Service Selection) contest 2010
  Retrieval Performance Evaluation of Matchmakers for Semantic Web Services.
  TC: OWLS-TC (version 4) and SAWSDL-TC

B. General announcement of the SEALS Evaluation Campaigns

In order to announce the various SEALS Evaluation Campaigns, the E.C.O.C. will distribute a general communication using any mechanism available (e.g., mailing lists, blogs, etc.) with the goal of reaching the developers of the tools that are targeted by the evaluation scenarios. This appendix contains this announcement and incorporates short explanations of each area's campaign and targets the general semantic web community.

Dear Sir/Madam,

We are pleased to announce that the first of the two public world-wide SEALS Evaluation Campaigns for semantic technologies will take place during the summer of 2010. The SEALS project will create a lasting reference infrastructure for automated semantic technology evaluation (the SEALS Platform) and thus facilitate the continuous evaluation of semantic technologies at a large scale with minimal effort from participants.

We cordially invite you to participate in the SEALS Evaluation Campaign in one or more of the five core areas shown below. Participation is open to anyone who is interesting in benchmarking a semantic technology tool. Detailed information regarding each area's campaign together with terms and conditions and general information about SEALS can be found on the SEALS Portal at http://www.seals-project.eu.

SEALS evaluations will address five core semantic technology areas:

Ontology Engineering Tools: addresses the ontology management capabilities of semantic technologies in terms of their ontology language conformance, interoperability and scalability. The main tools targeted are ontology engineering tools and ontology management frameworks and APIs; nevertheless, the evaluation is open to any other type of semantic technology.

Storage and Reasoning Systems: assesses a reasoner's performance in various scenarios resembling real-world applications. In particular, their effectiveness (comparison with pre-established 'golden standards'), interoperability (compliance with standards) and scalability are evaluated with ontologies of varying size and complexity.

Matching Tools: builds on previous matching evaluation initiatives (OAEI campaigns) and integrates the following evaluation criteria: (a) conformance with expected results (precision, recall and generalizations); (b) performance in terms of memory consumption and execution time; (c) interoperability, measuring the conformance with standard such as RDF/OWL; and (d) measuring the coherence of the generated alignments.

Semantic Search Tools: evaluated according to a number of different criteria including query expressiveness (means by which queries are formulated within the tool) and scalability. Given the interactive nature of semantic search tools, a core interest in this evaluation is the usability of a particular tool (effectiveness, efficiency, satisfac-

tion).

Semantic Web Service Tools: focuses on activities such as discovery, ranking and selection. In the context of SEALS, we view a SWS tool as a collection of components (platform services) of the Semantic Execution Environment Reference Architecture (SEE-RA). Therefore, we require that SWS tools implement one or more SEE APIs in order to be evaluated.

C. Long announcements of evaluation campaigns

Once the different evaluation scenarios are defined, the E.C.O.C. must announce the evaluation campaign using any mechanism available (e.g., mailing lists, blogs, etc.) with the goal of reaching the developers of the tools that are targeted by the evaluation scenarios. This appendix contains the full announcements of each area's evaluation campaign. These will target tool providers and researchers in specific areas.

C.1 Common to all tool areas

C.1.1 Instructions on how to participate

Tool developers are cordially invited to participate in the SEALS Evaluation Campaign in one or more of the five core areas. Participation is open to developers interested in evaluating their tool or to anyone who wants to evaluate a certain tool.

From the tool provider's perspective, the SEALS Evaluation Campaign has been designed to require minimal effort on their part. For the majority of SEALS Evaluation Campaigns, the evaluation process is fully automated from the point at which the tool has been uploaded to the SEALS Platform. The SEALS Platform, which will be available late July 2010, will manage all aspects of the evaluation. Once the evaluation has been completed, the results and analyses will be available from the SEALS Portal. In order to facilitate interaction between your tool and the SEALS Platform, the SEALS teams from each technology area have defined an easy to use API. Details about each technology area's API is available from their SEALS Evaluation Campaign's Portal page (http://www.seals-project.eu).

C.1.2 Evaluation Process

1. Preparatory Phase (March 2010 – July 2010)
   The first stage of this phase is the release of the detailed SEALS Evaluation Campaign materials. These will consist of

   (a) Instructions on how to participate.
   (b) The evaluation description.
   (c) The evaluation test data.
   (d) Any software needed for the evaluation.

   Once these materials have been made available and the SEALS Platform has been released, participants will be able to submit and install their tools on the SEALS platform. During this phase, the tool provider assisted by the E.C.E.C. have to verify that the tools submitted for evaluation run correctly on the SEALS platform. In addition, first tests can be run with datasets associated with that SEALS Evaluation Campaign.

2. Execution Phase (August – September 2010)
   Participants can submit different versions of their tool and use the SEALS plat-

form to evaluate them at any moment. When they feel confident that it is correctly configured, they can submit the final version to the system. It will then be evaluated and the raw results of the evaluation experiments will be stored in the SEALS repository.

3. Evaluation Phase (September – October 2010)
   The evaluation results will be automatically generated by the SEALS platform using the evaluation criteria identified by your tool's technology area E.C.E.C.. This process will generate different data representations, such as graphs and tables for visualizing recall/precision, time execution, and comparisons between the different campaign participants. The results will be publicly accessible in the SEALS platform after the evaluation phase has been conducted.

C.1.3 How to find out more

Detailed information regarding the SEALS Evaluation Campaign together with terms and conditions and general information about SEALS can be found on the SEALS Portal at http://www.seals-project.eu.

Each technology area's SEALS Evaluation Campaign is organised and executed by a SEALS *Evaluation Campaign Executing Committee* (E.C.E.C.). For more information on the SEALS Evaluation Campaign for this specific technology area, please contact the E.C.E.C.:
INSERT YOUR E.C.E.C. TABLE HERE

C.2 Ontology Engineering Tools

When dealing with the many existing semantic technologies, some questions arise sooner or later:

- Tool A is able of managing OWL DL ontologies but, up to what extent can it manage OWL Full ontologies?
- I am using an OWL Full ontology in Tool B and I want to use it in Tool C, which only supports OWL Lite. Can I make it with a minimal loss of information?
- Someone recommended me Tool D, but I need to manage very big ontologies. Can this tool make it efficiently? If not, which one can?

The SEALS Yardsticks For Ontology Management is an evaluation campaign that comprises a set of evaluations defined with the goal of evaluating the ontology management capabilities of semantic technologies in order to answer those questions.

The main tools targeted for these evaluations are ontology engineering tools and ontology management frameworks and APIs; nevertheless, the evaluation is open to any other type of semantic technology.

The evaluations will cover the evaluation of three characteristics of semantic technologies: conformance, interoperability and scalability

- Conformance. We will evaluate tool conformance with regards to the RDF(S) and OWL ontology languages with the goal of analysing up to what extent the

different ontology constructors are supported by tools. To this end, we will use four different test suites to cover the RDF(S), OWL Lite, OWL DL, and OWL Full languages.

- Interoperability. We will evaluate the interoperability of tools when interchanging ontologies using an interchange language with the goal of knowing the effects of interchanging ontologies between tools. As in the conformance evaluation, we will cover RDF(S), OWL Lite, OWL DL, and OWL Full as interchange languages.
- Scalability. We will evaluate the scalability of tools when managing ontologies of increasing size with the goal of checking up to what extent tools are able of dealing with big ontologies while maintaining their efficiency. In the scalability evaluation we will use synthetically generated ontologies of increasing size.

The evaluation campaign will take place during the summer of 2010.

Participation is open to developers interested in evaluating their tool or to anyone who wants to evaluate a certain tool.

Participants are just expected to collaborate in the connection of their tool with the SEALS Platform, which will be the infrastructure that will run all the evaluations automatically. Besides checking their results and comparing with others, once the tool is connected to the SEALS Platform participants will also be able to run the evaluations on their own with these and future test data.

If you want to participate, register your tool in the SEALS Portal (http://www.seals-project.eu) and stay tuned to the evaluation campaign web page where you can find detailed descriptions of the evaluations that we will perform and the latest information and results of the evaluation campaign.

This evaluation campaign is taking place inside the SEALS project. Go to the SEALS web page (http://www.seals-project.eu) and check the other evaluation campaigns that are taking place this year.

If you have any question or comment about the evaluation campaign, please contact us.

We count on your participation!

C.3 Storage and Reasoning Systems

Semantic technologies are at the heart of the future Web providing ways to express knowledge and data so that it can be properly exploited. These technologies will empower a new class of Information and Communication Technologies much more scalable, interoperable, and with a higher degree of process automation support that will fulfill the needs of an emergence market that will exceed $10 billion by 2010.

Description logic based systems (DLBSs) evaluation aims at accessing interoperability and performance of one of the core building blocks of semantic technologies vision namely description logic reasoners. This evaluation, thus, serves several purposes:

- Helps researches and users to select appropriate technologies;
- Gives an objective comparison among existing technologies;

- Stimulates technologies advancement through continuous evaluation.

The overall objective of evaluation campaign is to evaluate DLBSs standard inference services: classification, class satisfiability, ontology satisfiability and logical entailment. The challenge uses a set of set of state of the art ontologies for evaluation. The set includes OWL 2 test cases repository, ontologies from Gardiner suite, various versions of the GALEN ontology, ontologies that have been created in EU funded projects SEMINTEC, VICODI, AEO and SNOMED CT ontology. DLBSs are expected to support OWL 2 language and provide interface to their functionalities through a set of evaluation interfaces defined in the SEALS evaluation platform http://www.seals-project.eu/news/2-seals-platform.

C.4 Matching Tools

C.4.1 Evaluation Description

The first SEALS Evaluation Campaign aims at evaluating the competence of matching systems with respect to different evaluation criteria. The evaluation will focus on demonstrating the feasibility and benefits of automating matching evaluation:

- comparing two ontologies written in the same language: OWL DL,
- without input alignment,
- with any kind of fixed parameters and any kind of fixed and general purpose resources,
- without any kind of user input nor training samples.

Assumptions for this first campaign are that the matching systems can run independently and that it is possible and useful to compare systems based on different criteria separately. A limited set of criteria will be considered in this first evaluation:

- Conformance: standard precision and recall, restricted semantic precision and recall, coherence.
- Efficiency: runtime, memory consumption;
- Interoperability: compliance to the standard language RDFS and OWL DL;

C.4.2 Evaluation Test Data

We have selected a subset of the datasets that have been involved in previous OAEI campaigns. The datasets were selected based on the existence of reliable reference alignments and experiences with using the datasets in evaluation campaigns. These criteria are met by the following datasets:

- Systematic Benchmark: the goal of this benchmark series is to identify the areas in which each matching algorithm is strong or weak. The test is based on one particular ontology dedicated to the very narrow domain of bibliography and a number of alternative ontologies on the same domain for which alignments are provided.

- Conference: collection of conference organization ontologies. Compared to the benchmark series, the conference collections consists of heterogeneous conceptualisations that differ to a large degree. We will also measure the coherence of the generated alignments.
- Anatomy: the anatomy real world case is about matching the Adult Mouse Anatomy and the NCI Thesaurus describing the human anatomy. Compared to the other datasets it covers a very specific domain using a specialized vocabulary. Due to the large size of the ontologies it will also be used for analyzing runtime/memory efficiency.

In addition to the information above regarding how to participate in a SEALS campaign, tools submitted to the Matching Evaluation Campaign should be wrapped in a minimal standard interface (that of the Alignment API). In its minimal form, this amounts to implementing one class and one method of this class. Directions for achieving this are given at http://alignapi.gforge.inria.fr/tutorial/tutorial3/. The submission of the tool will follow the SEALS packaging procedure.

C.5 Semantic Search Tools

C.5.1 Introduction

The goals of the semantic search tool evaluation initiative are to support developers to improve their tools; compare their tools against their competitors and to generally improve the interoperability of semantic technologies.

The short-term goal is to create a set of reference benchmark tests for assessing the strengths and weaknesses of the available tools and to compare them with each other. As such, these tests will focus on the performance of fundamental aspects of the tool in a strictly controlled environment / scenario rather than their ability to solve open-ended, real-life problems.

C.5.2 Criteria

For the first evaluation campaign semantic search tools will be evaluated according to a number of different criteria including query expressiveness, usability (effectiveness, efficiency, satisfaction) and scalability. Scalability will address a number of factors including the tool's ability to query a large repository in a reasonable time; the tool's ability to cope with differing ontology sizes; and the tool's ability to cope with a large amount of query results. Query expressiveness will investigate the means by which queries are formulated within the tool and the degree to which this facilitates (or even impedes) the user's question-answering goal. However, given the interactive nature of semantic search tools, a core interest in this evaluation is the usability of a particular tool.

C.5.3 Two phase approach

The core functionality of a semantic search tool is to allow a user to discover one or more facts or documents by inputting some form of a query. The manner, in which this input occurs (e.g.: natural language, keywords, visual representation) is not of concern; however, the user experience of using the interface is of interest. Therefore, it is essential, that the evaluation procedures described in this document emphasize the users' experience with each tool.

In order to achieve this goal, the evaluation of each tool is split into two complementary phases: the automated phase and the user-in-the-loop phase. The user-in-the-loop phase involves a series of experiments involving human subjects, who are given a number of tasks (questions) to complete using a particular tool operating on an particular ontology.

Hence, the two core implications of this are that the user-in-the-loop experiments will be run by each tool provider participating in the evaluation and that additional software will be provided by this workpackage in order to both run the experimental workflows and likewise obtain the test data and return the results data to the various SEALS repositories. All materials required for the user-in-the-loop experiments will be provided by the SEALS consortium.

C.5.4 Evaluation Test Data

Only OWL ontologies will be used as test data: in order to simplify the development of the benchmarks for the first evaluation campaign, it has been decided that search tools operating on purely OWL ontologies will be evaluated. The evaluation of tools operating over a wider set of resources, e.g., OWL ontologies and document repositories, will be considered for the second evaluation campaign in late 2011 / early 2012.

For the first campaign, we have selected two datasets (one per evaluation phase).

- EvoOnt is a set of software ontologies and data exchange format based on OWL. It is well suited to the automated phase, since the A-Box is able to scale easily. It is, therefore, possible to produce various sizes of the same base ontology in order to measure the scalability and performance of a tool.

- The Mooney Natural Language Learning Data Set - has already been used in several evaluations before. The Mooney data comprises three data sets each supplying a knowledge base, English questions, and corresponding logical queries. The availability of the questions and groundtruths make it appropriate to the user-in-the-loop phase.

C.6 Semantic Web Service Tools

The SWS discovery activity consists of finding Web Services based on their semantic descriptions.

In this evaluation campaign scenario we will evaluate the retrieval and execution time performance of SWS tools for SWS discovery. The goal is to provide a forum and

infrastructure for the automatic evaluation of the performance of publicly available SWS tools over given public test collections.

The participating tools for SWS discovery or matchmaking will use formal semantics of both Goals (requests), i.e. the description of what the user is looking for; and Services (offers), i.e. the description of what a service does, written in the same description language or service ontology.

SWS discovery evaluation follows on the same principles and techniques from the more established Information Retrieval (IR) evaluation research area. Therefore we will use some common terminology and refer to common measures such as Precision and Recall.

This evaluation scenario will require that participants use the SEALS platform to perform evaluations. Access to the platform, including repositories will be available shortly. In addition, participating tools are required to implement the SEE API, to be downloaded from here.

The SEALS repositories will be used to register relevant information about participants and their tools.

Metadata about the API implementation should include:

- Implemented plugin - SEE API
- Implemented SEE services - e.g. discovery, storage
- Implemented operations - e.g. retrieveOntology(), retrieveServiceDescription(), retrieveGoalDescription(), discover(), select(), rank()

Metadata about the test data should include:

- Supported domain ontology language - e.g. WSML, OWL
- Supported SWS ontology language - e.g. WSMO, WSMO-Lite, OWL-S
- Ontology location (URL)
- Ontology name (URI, namespace)

Metadata about the tool should include:

- Name of tool/system
- Implemented campaign scenario
- Submission date
- Name and affiliation of authors
- Contact person and e-mail
- Related website
- Related papers
- URL to access installation package or Web page of the system

D. Evaluation campaign agreements

The following agreements describe the general terms for participation in the SEALS Evaluation Campaigns and the policies for using the resources and results produced in these evaluation campaigns. These were defined in SEALS Deliverable 3.1 (García-Castro and Martín-Recuerda, 2009) and are reproduced here for completeness.

D.1 Terms of participation

By submitting a tool and/or its results to a SEALS Evaluation Campaign the participants grant their permission for the publication of the tool results on the SEALS web site and for their use for scientific purposes (e.g., as a basis for experiments).

In return, it is expected that the provenance of these results is correctly and duly acknowledged.

D.2 Use rights

In order to avoid any inadequate use of the data provided by the SEALS Evaluation Campaigns, we make clear the following rules of use of these data.

It is the responsibility of the user of the data to ensure that the authors of the results are properly acknowledged, unless these data are used in an anonymous aggregated way. In the case of participant results, an appropriate acknowledgement is the mention of this participant and a citation of a paper from the participants (e.g., the paper detailing their participation). The specific conditions under which the results have been produced should not be misrepresented (an explicit link to their source in the SEALS web site should be made).

These rules apply to any publication mentioning these results. In addition, specific rules below also apply to particular types of use of the data.

D.2.1 Rule applying to the non-public use of the data

Anyone can freely use the evaluations, test data and evaluation results for evaluating and improving their tools and methods.

D.2.2 Rules applying to evaluation campaign participants

The participants of some evaluation campaign can publish the results as long as they cite the source of the evaluations and in which evaluation campaign they were obtained.

Participants can compare their results with other published results on the SEALS web site as long as they also:

- compare with the results of all the participants of the same evaluation scenario; and
- compare with all the test data of this evaluation scenario.

Of course, participants can mention their participation in the evaluation campaign.

D.2.3 Rules applying to people who did not participate in an evaluation campaign

People who did not participate in an evaluation campaign can publish their results as long as they cite the sources of the evaluations and in which evaluation campaign they were obtained and they need to make clear that they did not participate in the official evaluation campaign.

They can compare their results with other published results on the SEALS web site as long as they:

- cite the source of the evaluations and in which evaluation campaign they were obtained;
- compare with the results of all the participants of the same evaluation scenario; and
- compare with all the test data of this evaluation scenario.

They cannot pretend having executed the evaluation in the same conditions as the participants. Furthermore, given that evaluation results change over time, it is not ethical to compare one tool against old results; one should always make comparisons with the state of the art.

D.2.4 Rules applying to other cases

Anyone can mention the evaluations and evaluation campaigns for discussing them.

Any other use of these evaluations and their results is not authorized (you can ask for permission however to the contact point) and failing to comply to the requirements above is considered as unethical.